

Language Models are Realistic Tabular Data Generators

Year: 2022 | Citations: 321 | Authors: V. Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, Gjergji Kasneci

Abstract

Tabular data is among the oldest and most ubiquitous forms of data. However, the generation of synthetic samples with the original data's characteristics remains a significant challenge for tabular data. While many generative models from the computer vision domain, such as variational autoencoders or generative adversarial networks, have been adapted for tabular data generation, less research has been directed towards recent transformer-based large language models (LLMs), which are also generative in nature. To this end, we propose GReaT (Generation of Realistic Tabular data), which exploits an auto-regressive generative LLM to sample synthetic and yet highly realistic tabular data. Furthermore, GReaT can model tabular data distributions by conditioning on any subset of features; the remaining features are sampled without additional overhead. We demonstrate the effectiveness of the proposed approach in a series of experiments that quantify the validity and quality of the produced data samples from multiple angles. We find that GReaT maintains state-of-the-art performance across numerous real-world and synthetic data sets with heterogeneous feature types coming in various sizes.